APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Jan. 2005, p. 80–84 Vol. 71, No. 1
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.1.80–84.2005

Use of Artificial Neural Networks To Accurately Identify *Cryptosporidium* Oocyst and *Giardia* Cyst Images

Kenneth W. Widmer, Deepak Srikumar, and Suresh D. Pillai*

Food Safety and Environmental Microbiology Program, Poultry Science Department, Institute of Food Science and Engineering, Texas A&M University, College Station, Texas

Received 19 May 2004/Accepted 6 August 2004

***Cryptosporidium parvum* and *Giardia lamblia* are protozoa capable of causing gastrointestinal diseases. Currently, these organisms are identified using immunofluorescent antibody (IFA)-based microscopy, and identification requires trained individuals for final confirmation. Since artificial neural networks (ANN) can provide an automated means of identification, thereby reducing human errors related to misidentification, ANN were developed to identify *Cryptosporidium* oocyst and *Giardia* cyst images. Digitized images of *C. parvum* oocysts and *G. lamblia* cysts stained with various commercial IFA reagents were used as positive controls. The images were captured using a color digital camera at 400× (total magnification), processed, and converted into a binary numerical array. A variety of "negative" images were also captured and processed. The ANN were developed using these images and a rigorous training and testing protocol. The *Cryptosporidium* oocyst ANN were trained with 1,586 images, while *Giardia* cyst ANN were trained with 2,431 images. After training, the best-performing ANN were selected based on an initial testing performance against 100 images (50 positive and 50 negative images). The networks were validated against previously "unseen" images of 500 *Cryptosporidium* oocysts (250 positive, 250 negative) and 282 *Giardia* cysts (232 positive, 50 negative). The selected ANNs correctly identified 91.8 and 99.6% of the *Cryptosporidium* oocyst and *Giardia* cyst images, respectively. These results indicate that ANN technology can be an alternate to having trained personnel for detecting these pathogens and can be a boon to underdeveloped regions of the world where there is a chronic shortage of adequately skilled individuals to detect these pathogens.**

Cryptosporidium parvum and *Giardia lamblia* are intestinal parasites capable of infecting humans and causing fatal or life-threatening gastroenteritis (8, 12). Drinking water contamination by these pathogens is a serious concern, and *C. parvum* is currently regulated by the U.S. Environmental Protection Agency (USEPA) by the Long-Term 2 Treatment Rule (14). Monitoring for these protozoa in drinking water, source water, effluent, and foods is of significant public health importance (4, 6, 7, 10).

The USEPA-recommended method for the detection of oocysts and cysts is immunomagnetic separation and capture, followed by immunofluorescent antibody (IFA) staining and fluorescence microscopic confirmation based on morphological characteristics (13). Microscopic interpretation of the IFA-stained cysts and oocysts is a key step in the monitoring of *Cryptosporidium* oocysts and *Giardia* cysts. USEPA method 1623 requires technically proficient analysts for final confirmation (13). The identification of *Cryptosporidium* and *Giardia* contamination is thus totally dependent upon the experience of the analyst. Human error, however, does and can contribute significantly to the problems with identification (2, 9).

Artificial neural networks (ANN) implement algorithms to mimic the neuron processing functions of true neural networks, where neurons set in layers (each neuron being connected to all other neurons in the preceding layer) process information. The information is applied to an activation function and (if reaching a specific threshold) passes an output signal to other neurons within the system. Using this framework of interconnected neurons, neural networks undergo a training procedure where they "learn" to discern patterns in data (1, 17).

The back-propagation algorithm is a common learning algorithm employed by ANN. This learning process involves two steps, the first step being a forward processing of input data by the neurons that produces a predicted solution. The second step is an adjustment of weights within the neuron layers (sequentially from the outputs back through the network) in order to minimize the errors of the predicted solution compared to the true (correct) solution. Another set of input data is presented and the process is repeated, eventually reducing the amount of errors by continual weight adjustment so that the predicted output matches the true output (17).

ANNs have been used successfully as image classification systems for plankton (3, 11), inflammatory cells (18), and cervical neoplasia smears (5). We have previously proven the concept that ANN can be used to detect *C. parvum* oocysts (15, 16). The objective of the present study was to expand the technology to include the other key protozoan (*G. lamblia*) and to exploit the image data features, such as shading characteristics and precise shape. The ability to incorporate these image attributes into the ANN is a significant improvement over previous approaches.

* Corresponding author. Mailing address: 418D Kleberg Center, TAMUS 2472, Texas A&M University, College Station, TX 77843-2472. Phone: (979) 845-2994. Fax: (979) 845-1921. E-mail: spillai@poultry.tamu.edu.

MATERIALS AND METHODS

***C. parvum* and *G. lamblia* (positive) samples and image processing.** *C. parvum* and *G. lamblia* oocyst and cyst samples used in the training and testing of the

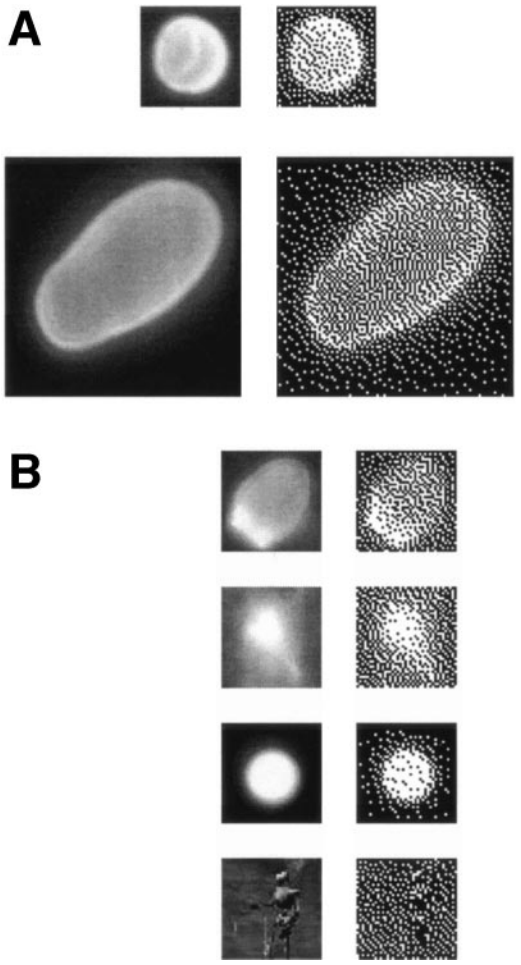

FIG. 1. (A) Images of a *Cryptosporidium* oocyst (top) and *Giardia* cyst (bottom) before and after dither filtering. (B) Examples of negative images after dither filtering. From top to bottom: algal samples, surface sediment debris, microspheres, and digital image artwork.

ANN were obtained from two commercial suppliers (Waterborne Inc., New Orleans, La.; Sterling Parasitology Lab, Tucson, Ariz.). Portions of the parasite stocks were mounted on well slides and stained with commercially available, fluorescein-labeled monoclonal antibodies following the manufacturer's protocol (AquaGlo from Waterborne Inc.; Crypto/Giardia IF test from TechLab, Blacksburg, Va., and Meridian Bioscience Inc., Cincinnati, Ohio). The slides were stored at 4°C in the dark until microscopic observation. All slides were observed at 400× total magnification using either a BH-2 Olympus, an Auxioplan 2 Zeiss, or a BX-50 Olympus microscope. The samples were observed under fluorescence illumination generated by an attached 100-W mercury lamp possessing filters for UV excitation (398 or 490 nm).

The fluorescent microscopic images were captured using a charge-coupled device (CCD) color digital camera (SPOT CCD; catalog no. SP100; Diagnostic Instruments, Inc., Sterling Heights, Mich.). The images were collected over a 5-year period (January 1998 to June 2003). All images were saved in a color TIFF format by using a PC system.

Individual oocyst and cyst images were cropped (40 by 40 and 95 by 95 pixels, respectively) from their original size by using commercial software (Adobe PhotoShop 5.5; Adobe Systems Incorporated, San Jose, Calif.). The color information of these images was discarded by utilizing the dither filter option of another commercial software (Xnview 1.168) to convert these images into black-and-white RAW files (Fig. 1). Using a C++ program, we converted the black-and-white images into binary numerical arrays (of either 40 by 40 or 95 by 95 elements). Each array value represented the corresponding pixel color of "black" or "white" from the original cropped RAW image file.

Non-*C. parvum* and non-*G. lamblia* negative images. Negative images were used for ANN training and testing. These images consisted of cross-reacting algal samples, green fluorescent spheres, environmental matrices, or digital image artwork. Additionally, fluorescence microscopic images of a surface water sediment concentrate were also obtained as mentioned above. All negative images were cropped and processed in a manner similar to that for the positive images (Fig. 1B).

Network image file generation for testing and training. The *Giardia* and *Cryptosporidium* images were separated into image sets by using a text editor. The two sets of files were either positive images (*C. parvum* oocysts or *G. lamblia* cysts) or negative images (non-*C. parvum* or non-*G. lamblia*). To reduce bias, each of the files had the order of the images randomized by using a utility function from the ANN program (BrainMaker Professional; California Scientific Software, Nevada City, Calif.). Images were randomly selected for the network testing sets, and the remaining images were used for network training.

For the oocyst networks, the training set consisted of 1,586 images (774 positive and 812 negative). The cyst network training set consisted of 2,431 images (1,521 positive and 910 negative images). For the cyst network, the number of unique negative images was much less than the number of positive images (only 182 images compared to 1,521). So, these images were replicated five times to increase the number of negative images presented to the ANN during training. This was necessary to prevent poor training from an extremely unbalanced image data set (positive compared to negative images).

Two types of testing files were generated for each network type: the initial testing and the validation testing files. The initial testing files for each type of network (cyst or oocyst) consisted of 100 images (50 positive images and 50 negative images). This initial testing was performed initially to identify which network(s) performed well (in terms of percent correct identification). The validation testing files had either 500 oocyst images (250 positive and 250 negative) or 282 cyst images (232 positive and 50 negative). These image sets were used to validate the best-performing ANN.

Network training and testing. The ANN were developed using a commercial software program (BrainMaker Professional; California Scientific Software) which utilized a back-propagation algorithm and was run on a PC system. Each network type had different numbers of input neurons. The oocyst networks had 1,600 input neurons, while the cyst networks had 9,025 input neurons. Both ANN designs had five hidden and two output neurons. Each output had values ranging from 0 to 1, dependent on the image being classified as positive or negative.

During training, the results of the initial testing performances were compiled. The networks were trained for 150 runs (all training images were presented during each training run). After each training run, the generated networks were saved and tested against the initial testing set. As networks were tested, no adjustments were made to alter decision-making nodes. An image was scored as correct or incorrect after comparing the predicted output values to the image's true output values. A correct identification was an output value of 0.900 or higher, while any other result was considered an incorrect identification (e.g., a test image having output values of 0.994 and 0.004 would be classified as a positive image which had true output values of 1 and 0).

From these initial testing results, three networks (which identified the most images correctly) were selected for further testing against the validation testing image sets. For each network that was tested, its performance was recorded as a percentage of correct identifications for the primary and validation image sets.

RESULTS

The networks designed to identify *C. parvum* oocysts tested well with the initial testing set. From the initial testing (100 testing images), three selected networks (runs 61, 125, and 126) identified 93 of these images correctly (Fig. 2). For the validation testing (500 images), the correct identification percentages ranged from 88.6% (run 126) to 91.8% (run 61) (Fig. 2). The *Giardia* cyst ANN trained very well, producing several networks capable of identifying the training images with great success. Runs 2, 3, and 4 were selected for further testing, based on their performance against the initial testing set for which these networks identified 89 to 95 of the 100 initial testing images correctly (Fig. 3). During validation, the percent correct identification ranged from 98.9% (run 2) to 99.6% (run 4) (Fig. 3).

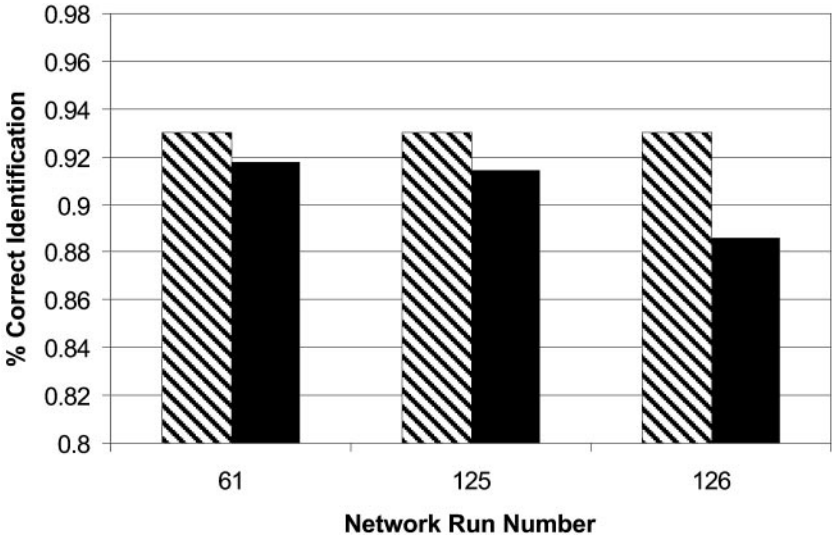

FIG. 2. Testing results for *Cryptosporidium* oocyst networks. Cross-hatched bars represent the initial testing results from 100 images (50 positive and 50 negative); solid bars represent the validation testing results of 500 images (250 positive and 250 negative).

It is critical that ANN results be also analyzed to determine the percentage of false positives and false negatives. The results based on the misclassified images from the validation tests of the oocyst and cyst networks are summarized in Table 1. There appears to be a general bias towards false positives (misclassified negatives, i.e., incorrectly identified negative images as oocysts or cysts). The *Giardia* ANN exhibited a lower occurrence of misclassified images. There were no instances where these networks incorrectly classified a true positive image as negative. The false positive identification in the *Giardia* ANN ranged between 2% (1 of 50) in run 4 and 6% (3 of 50) in run 2.

There was a slightly higher misidentification rate with the *Cryptosporidium* ANN compared to the *Giardia* ANN. The false positives (misclassifying negatives as positives) in the *Cryptosporidium* ANN ranged between 12% (30 of 250) and 14% (35 of 250). The false negatives (misclassified positives) in the *Cryptosporidium* network, in contrast, ranged between 4.4% (11 of 250) in run 61 and 8.8% (22 of 250) in run 126.

DISCUSSION

The three networks that were chosen for validation were based on results obtained from the initial testing runs. The *Giardia* ANN and *Cryptosporidium* ANN were both initially tested using 100 images. There is no apparent relationship between the validation results and the proportion of initial testing images to training images. For the *Giardia* ANN it was 4.1% (100 of 2,431), and for the *Cryptosporidium* ANN it was 6.3% (100 of 1,586). The percentages of correct identification

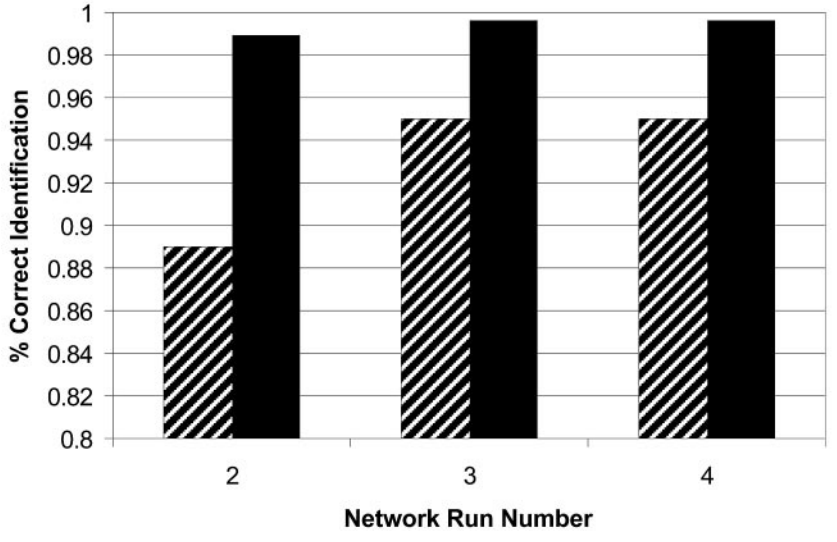

FIG. 3. Testing results for *Giardia* cyst networks. Cross-hatched bars represent initial testing results for 100 images (50 positive and 50 negative); solid bars represent the validation testing results of 282 images (232 positive and 50 negative).

TABLE 1. Occurrence of misclassified images based on validation test results

Network and run no.	% Misclassified pos.[a]	% Misclassified neg.[a]
Oocyst networks		
61	4.4 (11/250)	12 (30/250)
125	5.2 (13/250)	12 (30/250)
126	8.8 (22/250)	14 (35/250)
Cyst networks		
2	0 (0/232)	6 (3/50)
3	0 (0/232)	2 (1/50)
4	0 (0/232)	2 (1/50)

[a] Numbers in parentheses are the numbers of images misidentified out of the total in the validation testing set.

of the protozoa images were impressive. The *Giardia* ANN was upwards of 99.6% correct, while the *Cryptosporidium* ANN was 91.8%. (These results should be evaluated in comparison to the time and accuracy requirements to view the same 500 and 282 image sets). The current USEPA method considers a counting error rate of 10% to be acceptable variation from analyst to analyst (13). The results in this study indicate current ANN performance may be sufficient despite false-positive and false-negative identifications of oocysts ranging from 4.4 to 14% (Table 1).

It should be noted that the validation images used in this study are field realistic, in that they were prepared using a variety of IFA stains, which in reality would be the scenario if this technology were used to identify images obtained from different laboratories. The percentages of correct identifications, in spite of the inherent variation that occurs in FA staining, are an indication of the robustness of these ANN. It needs to be reiterated that all testing images (primary and validation) were unique in that these images were never presented to the networks during the training.

This work is supported by previous work by us utilizing neural networks as protozoa image classifiers (15, 16). In these previous studies the ANN performed comparably well but were limited in the number and variety of testing images utilized. Additionally, the present work entails an improvement from the image processing techniques that were employed in the previous studies, which were based on grayscale histograms (15) or summation of pixel values from binary threshold images (16). Zheng et al. (18) have used a similar image processing approach (i.e., pixel intensity variation) for ANN analysis of histological specimen images and have achieved 97 and 98% recognition rates. The numbers of images used in this study in the training and testing phases were larger than those used by Zheng et al. The image processing technique used in this study incorporates the entire shape and shading characteristics of the image (oocyst and cyst) of interest (Fig. 1A). Even though the images are black and white, the dither filtering still maintains some form of shading to the original image and, thus, the entire image, positional data, and shape may be used as information for decision making by the networks.

A salient feature of the ANN output is that it does not produce a single result (yes or no) but instead a range of numbers associated with two possible classifications. Because of this, a user has additional information that could help in identifying misclassified images. These ANN could be used in conjunction with our previously developed ANN to identify 4′,6-diamidino-2-phenylindole-stained *C. parvum* oocysts (15). An ANN could theoretically be developed to identify differential interference contrast images of cysts and oocysts as a means of final confirmation.

An ideal scenario for implementing the ANN technology will be the trained ANN used as the primary means of identifying oocysts and cysts through automated means. A user could prepare a sample slide and have an automated motorized microscope stage and digital camera capture images of interest, and ANN could then be used to identify these images as being cysts or oocysts. The time required for identifying hundreds of such images with a validated ANN is a few seconds. As a cost issue for an automated ANN system, primary expenses would stem from the motorized stage and computer-controlled digital camera. The ANN developed in this study do not require an advanced PC system to operate (and could be run on the same PC system used for controlling the digital camera).

The use of such ANN for the identification *Giardia* and *Cryptosporidium* images has several possible applications, especially with regards to water quality monitoring in underdeveloped countries. The system could be used as a means for identifying oocysts and cysts remotely, as images could be transmitted via the internet to a central location for identification. Alternatively, a central identification facility could use the ANN to achieve high-throughput identification of target organisms from slides shipped in from different locations. The ANN could also be used for initial classification of images on site, while a human analyst could be used to visually confirm the identifications remotely (looking at the digitally captured images). The main advantage of using ANN would be overcoming the need to have a trained analyst on site. Such a feature would be highly advantageous in developing and under-developed countries, where trained analysts may not readily be available or would be cost-prohibitive to retain over multiple years. Rough calculations suggest that the ANN could save as much as 80% in costs by a laboratory not having to retain trained individuals. Additionally, ANN could be used as a teaching tool to train analysts, or they could provide an analyst a second opinion. Both of these implementations may be of particular use in under-developed countries. The ANN image classification systems may also be developed and utilized for the identification of other protozoa, such as *Cyclospora*, *Microsporidia*, or *Toxoplasma gondii*. It may be possible to develop several networks that could work in concert to identify a suite of protozoa and other parasitic organisms.

As advances are made in molecular detection techniques, there will be less reliance on conventional detection methods such as microscopy. Consequently, there will be declining numbers of college courses in microscopy and thereby fewer individuals with the expertise to identify protozoa by microscopy. ANN such as those described here may end up serving as a repository for knowledge, archiving the ability to identify microscopic images of oocysts, cysts, and other protozoa. In this mode, ANN might then serve as a valuable teaching tool to train individuals in microscopically identifying emerging or reemerging protozoan and other microscopically discernible human pathogens. Future work will be aimed at validating this technology in a multilaboratory round-robin format.

ACKNOWLEDGMENTS

This work was supported by funds from the U.S. EPA-STAR program (R829009), USDA-CSREES program (2001-34461-10405), USDA-CSREES/IFAFS program (00-52102-9637), and a Hatch grant (H8708). Portions of this work were performed by K.W. when he was an ORISE Research Fellow at the Office of Ground Water and Drinking Water at the U.S. EPA, Cincinnati, Ohio.

REFERENCES

1. **Basheer, I. A., and M. Hajmeer.** 2000. Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods **43:**3–31.
2. **Clancy, J., W. Gollnitz, and Z. Tabib.** 1994. Commercial labs: how accurate are they? J. Am. Water Works Assoc. **86:**89–96.
3. **Culverhouse, P., R. Ellis, R. Simpson, R. Williams, R. Pierce, and J. Turner.** 1994. Automatic categorization of five species of Cymatocylis (Protozoa, Tintinnida) by artificial neural network. Mar. Ecol. Prog. Ser. **107:**273–280.
4. **Gómez-Couso, H., F. Freire-Santos, J. Martínez-Urtaza, O. García-Martín, and M. E. Ares-Mazás.** 2003. Contamination of bivalve mollusks by Cryptosporidium oocysts: the need for new quality control standards. Int. J. Food Microbiol. **87:**97–105.
5. **Kok, M. R., M. E. Boon, P. G. Schreiner-Kok, J. Hermans, D. E. Grobbee, and L. P. Kok.** 2001. Less medical intervention after sharp demarcation of grade 1–2 cervical intraepithelial neoplasia smears by neural network screening. Cancer Cytopathol. **93:**173–178.
6. **LeChevallier, M. W., W. D. Norton, and R. G. Lee.** 1991. Occurrence of *Giardia* and *Cryptosporidium* spp. in surface water supplies. Appl. Environ. Microbiol. **57:**2610–2616. (Erratum, **58:**780, 1992.)
7. **LeChevallier, M. W., and W. D. Norton.** 1995. *Giardia* and *Cryptosporidium* in raw and finished water. J. Am. Water Works Assoc. **87:**54–68.
8. **Marshall, M. M., D. Naumovitz, Y. Ortega, and C. R. Sterling.** 1997. Waterborne protozoan pathogens. Clin. Microbiol. Rev. **10:**67–85.
9. **McClellan, P.** 1998. Sydney water inquiry: final report. Publication no. 628.1099441. New South Wales Premier's Department, Water Inquiry Secretariat, Sydney, Australia.
10. **Robertson, L. J., G. S. Johannessen, B. K. Gjerde, and S. Loncarevic.** 2002. Microbial analysis of seed sprouts in Norway. Int. J. Food Microbiol. **75:** 119–126.
11. **Simpson, R., R. Williams, R. Ellis, and P. Culverhouse.** 1992. Biological pattern recognition by neural networks. Mar. Ecol. Prog. Ser. **79:**303–308.
12. **Tzipori, S.** 1985. *Cryptosporidium*: notes on epidemiology and pathogenesis. Parasitol. Today **1:**159–165.
13. **U.S. Environmental Protection Agency.** 2001. Method 1623: *Cryptosporidium* and *Giardia* in water by filtration/IMS/FA. Publication no. EPA/821/R-01/025. Office of Ground Water and Drinking Water, Washington, D.C.
14. **U.S. Environmental Protection Agency.** 2003. Long-term 2 treatment rule. Publication no. EPA/816/D-03/001. Office of Ground Water and Drinking Water, Washington, D.C.
15. **Widmer, K. W., K. H. Oshima, and S. D. Pillai.** 2002. Identification of *Cryptosporidium parvum* oocysts by an artificial neural network approach. Appl. Environ. Microbiol. **68:**1115–1121.
16. **Widmer, K. W.** 2003. Development of artificial neural networks capable of identifying *Cryptosporidium parvum* oocysts stained with 4′,6-diamidino-2-phenylindole. J. Rapid Methods Autom. Microbiol. **11:**97–110.
17. **Widrow, B.** 1990. 30 years of adaptive neural networks: perception, madaline, and backpropagation. Proc. IEEE **78:**1415–1441.
18. **Zheng, Q., B. K. Milthorpe, and A. S. Jones.** 2004. Direct neural network application for automated cell recognition. Cytometry **57:**1–9.